Volume 14 Number 16 1986 Nucleic Acids Research

Quantitative analysis of the relationship between nucleotide sequence and functional activity

Gary D.Stormo, Thomas D.Schneider and Larry Gold

Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309-0347, USA

Received 22 April 1986; Revised and Accepted 14 July 1986

ABSTRACT

Matrices can be used to evaluate sequences for functional activity. Multiple regression can solve for the matrix that gives the best fit between sequence evaluations and quantitative activities. This analysis shows that the best model for context effects on suppression by *su2* involves primarily the two nucleotides 3' to the amber codon, and that their contributions are independent and additive. Context effects on 2AP mutagenesis also involve the two nucleotides 3' to the 2AP insertion, but their effects are not independent. In a construct for producing β-galactosidase, the effects on translational yields of the tri-nucleotide 5' to the initiation codon are dependent on the entire triplet. Models based on these quantitative results are presented for each of the examples.

INTRODUCTION

Several recent papers have used matrices to evaluate nucleic acid sequences (1-5). The matrix contains a value for each possible base at each position in a site. The values corresponding to the bases in the sequence are added to give the evaluation. This method assigns a value to any sequence (see Figure 1a).

The different papers vary in their methods of assigning values to the elements of the matrix. In this paper we show how to use methods for solving simultaneous equations to find the matrix elements that give the best fit to a set of quantitative data. If a site covers n bases there are 4^n different sequence possibilities for any site. However, there are only $4n$ elements in the matrix. Quantitative data for only $(3n+1)$ sequences is sufficient to solve for the matrix elements that will evaluate all 4^n possible sequences (see Methods and Fig. 2).

The goodness of fit to experimental data is a measure of the appropriateness of the matrix. The matrix of Figure 1a evaluates sequences according to their mononucleotides at particular positions. It is possible that the important information for a site's activity is embodied in di-nucleotide, or higher, combinations of bases that are not additive. If so, a mono-nucleotide matrix will not be found that adequately matches the quantitative data; this serves as a test of what sequence information is

being recognized. More complex matrices can be constructed when the data are not well fit by the simple mono-nucleotide matrix.

We have chosen three sets of data to demonstrate the method. We first use data from papers by Miller and Albertini (6) and Bossi (7) on the context effects on nonsense suppression. This is a large collection of data for which the mechanism of the effect is not understood. We do an analysis of the data for *su2* since it shows the greatest range of context effects. We next analyze data on the context effects on mutagenesis from a paper by Coulondre *et al.* (8). Finally, we examine data about the effects that the three nucleotides preceding the initiation codon have on the efficiency of translation initiation (9).

METHODS

Sequence/Activity Relationship as Simultaneous Equations

The matrix evaluation method assigns a quantitative value to any sequence, given the matrix elements. When quantitative data are known for many sequences we can solve for the matrix elements that give the best fit between the sequences and those data. Multiple linear regression is a method for giving the least squares best fit to data of this type. We have used the Minitab (10) data analysis package on the University of Colorado's CDC Cyber 720 for all the analyses described. Sequence manipulations and encoding of sequences into vectors for the numerical analyses were done using the programs of the Delila system (11,12).

The number of independent variables is actually less than the number of matrix elements, as can be seen in Figure 1. The matrix of Figure 1b gives every sequence the same value as the matrix of Figure 1a, but it has four fewer variables. The value for each **T** has been set to zero and a constant term has been added. The constant is the value given to the sequence of all **T**s. The other elements are the differences between having a **T** at each position and each of the other bases. The total number of independent variables is three times the number of positions, plus one. This treatment is dummy encoding, often seen in statistical analyses (13).

Alternative Data Representations - Test of Independence

The matrices of Figure 1 evaluate sequences according to their mono-nucleotides. However, the information in the sequence that determines the activity may not be the sum of the mono-nucleotide information. That is, there may be information in the di-nucleotides that is absent from the mono-nucleotide sum; the contributions of the mono-nucleotides to the activity being measured may not be independent. We can make the matrix fit the appropriate form of the sequence information. For instance, we can use di-nucleotides instead of mono-nucleotides as the rows of the matrix if they contain important additional information. As long as the information

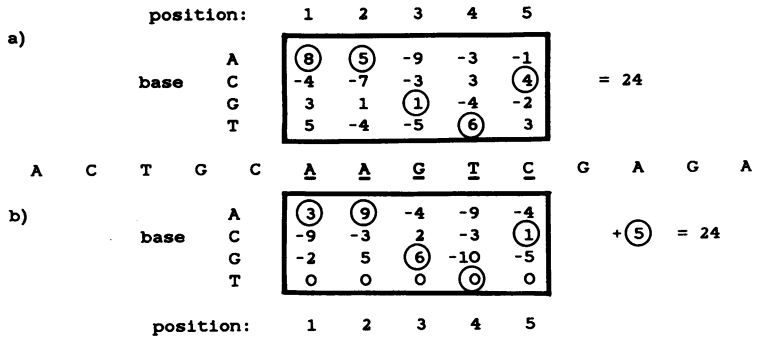


Figure 1. Matrix Evaluation of a Sequence. a) The matrix is used to evaluate any sequence by extracting and adding the element from each column that corresponds to the sequence (see [1]). The sequence *AAGTC* is evaluated to 24. b) This matrix has the elements for T's set to 0, and a constant (+5) is added. The constant and the remaining 15 matrix elements are the 16 variables solved for in the multiple regression. Every sequence is evaluated identically by this matrix and the one of part a.

content in the site is the sum of subsequences that are smaller than the entire site we gain some predictive value because the number of variables to solve for is less than the number of sites that can be evaluated with the matrix. Furthermore, the ability to find a good solution with multiple regression is a test of the appropriateness of the model. If mono-nucleotides give a good fit, there is no reason to invoke models involving higher units of information.

Logarithm of Activities

Typically the activities measured are the product of several partial reactions. The nucleotides at the interaction site contribute the partial reactions that combine to the overall interaction. If the equilibrium constant for each partial reaction is K_i, then the total equilibrium constant is

$$K_{total} = \prod_i K_i.$$

If the matrix elements are related to the partial reaction constants and they are to be added together to give a value related to the total activity, they must be proportional to the logarithm of the measured activity since

$$ln(K_{total}) = \sum_i ln(K_i).$$

That is, we solve for matrix elements that give a best fit to the logarithms of the measured activities. One can think of the matrix elements as being analogous to partial ΔG's which are added to give the total binding energy. The same treatment

can be applied to non-equilibrium reactions that are not too complicated, such as the $K_B k_2$ analysis of promoters (14).

RESULTS

Suppression of Amber Mutations by *su*2

Mono-nucleotides at Six Positions. Figure 2a summarizes the data for *su*2 suppression of various amber codons, from the papers of Albertini and Miller (6) and Bossi (7), for *E. coli* and *S. typhimurium*, respectively. (All of our sequences are displayed with T's for consistency; for RNA this represents *uridine*). The correlation between the activities measured in the two papers is 0.95, so we combined them into a single data set. We start with the assumption that the context effect is localized within the two codons surrounding the amber codon. Figure 2b displays the formal matrix whose elements we solve for, and Figure 2c displays a subset of the equations used in the multiple linear regression.

Figure 3a shows the matrix obtained by regression of the logarithms of the activities by those six mono-nucleotides. Figure 3b shows the plot of the observed values versus the values predicted from the matrix. $r^2 = 0.857$, indicating that most of the suppression activity can be accounted for by the addition of effects from neighboring mono-nucleotides. The matrix also shows that only at positions +3 and +4 is the suppression activity much affected by the different bases. At the other positions the identity of the base is largely irrelevant to the degree of suppression. Furthermore, the rules for suppression context are quite simple: a purine at position +3 is good for suppression, and a **T** at position +4 is also good. These two effects are nearly equal to each other, and the goodness of fit to the data (Fig. 3b) implies that they are independent of each other.

Mono-nucleotides at Two Positions. Figure 4 shows the matrix and plot where only the two positions following the amber codon are included in the analysis. The predictions are nearly as good as before ($r^2 = 0.765$), confirming that most of the suppression information is contained in these two positions.

Di-nucleotide at Positions +3 and +4. We also asked whether the two important positions are independent. Figure 5 shows the matrix and plot for the model in which the two following nucleotides are the only ones important, but the information is in the di-nucleotide. There are no matrix entries for the di-nucleotides **AA, GT, TA** or **TG** because these do not occur in the data set. The slight improvement ($r^2 = 0.876$) of this analysis over that in Figure 4 shows that most, but not all, of the context suppression effect is additive between the two bases.

Di-nucleotide at +3,+4 and Two Mono-nucleotides. Figure 6 shows the matrix and plot using a combination of information. The information of the di-nucleotide at

a)

amber sites	activities: (6)	(7)	amber context
O12	2.3	3.3	ATT tag TCT
A27	3.0	2.9	ACC " TCC
A9	2.5	1.1	AAA " TCG
A10	3.8	2.0	CAG " TTG
O17	0.8	0.8	GCG " CGC
X15	0.7	0.7	GAT " CGC
O25	1.8	1.0	GCG " CGG
O14	2.8	1.4	GTA " CGA
O18	1.7	3.4	AAC " CCG
A23	4.2	3.7	ATT " CCG
A19	3.8	3.8	CAC " CAA
O27	2.5	3.2	TTT " CAA
09	5.1	5.9	GCA " CAA
O13	11.3	11.2	GAT " CTG
O10	12.3	10.0	CAA " CTG
O36	10.3		GTC " CTG
A11	8.1	7.3	TCG " CTG
O35	7.5	5.0	CTG " CTC
O23	11.5	9.6	AAA " CTC
O15	8.3	7.2	GTC " GCC
A6	12.9	7.5	AAC " GCC
O19	9.3	7.8	GTG " GCT
A34	8.3	8.9	GGC " GCG
A20	12.3	9.4	GTC " GCG
A33	10.6	6.5	TCT " GGC
X3	18.4	19.7	CAA " GGT
A18	14.2	14.5	GCA " GGT
O30	10.6	10.9	GGA " GAC
A29	10.4		GGA " GAC
A15	13.8	14.0	GAC " GAT
A31	12.1	11.2	AAA " GAT
O11	18.0	17.7	TCG " ATT
O24	20.7	17.0	AAT " ATT
O21	14.6	16.5	CAG " ATC
A30	25.6	19.5	TGT " ATC
A5	9.4	6.4	TAT " ACC
O28	10.1	11.9	CAA " ACC
O34	9.2	8.6	GGG " ACC
A16	9.6	11.5	GAC " ACA
A24	10.8	10.8	GAC " AGT
A13	19.3	19.9	GTG " ATG
A26	17.8	16.8	GAT " ATG
A17		1.3	GTG " CAT

b)

x_1	x_4	x_7	tag	x_{10}	x_{13}	x_{16}
x_2	x_5	x_8	"	x_{11}	x_{14}	x_{17}
x_3	x_6	x_9	"	x_{12}	x_{15}	x_{18}
0	0	0	"	0	0	0

c)

$x_3 + x_6 + x_9 + x_{10} + x_{14} + x_{17} + C = ln(8.6)$

$x_3 + x_4 + x_8 + x_{10} + x_{14} + x_{16} + C = ln(11.5)$

$x_3 + x_4 + x_8 + x_{10} + x_{15} + 0 + C = ln(10.8)$

$x_3 + 0 + x_9 + x_{10} + 0 + x_{18} + C = ln(19.9)$

etc.

Figure 2. Supression Activities of Different Amber Contexts. a) Activites are from Miller and Albertini (6) and Bossi (7). The numbers are the percent activity of β-galactosidase when the amber codon is suppressed by *su2* as compared to the non-amber (wild-type) codon. Capital letter nucleotides are the variables in the analysis. b) This shows the matrix as a collection of unknowns that are to be determined from the simultaneous equations. c) A few of the simultaneous equations are displayed, using the data from Bossi (7). There is an equation for each of the measured activities. From these equations the unknowns x_1 to x_{18} and C can be determined.

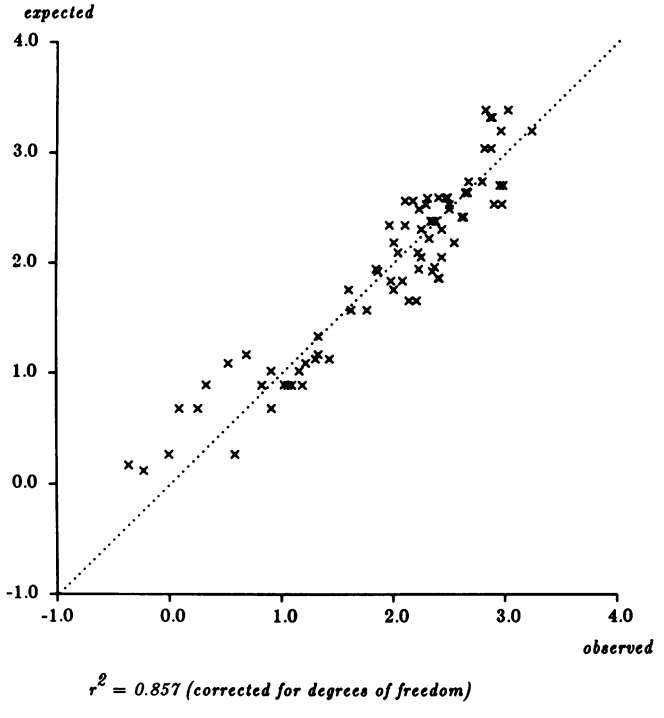
position +3,+4 is combined with the mono-nucleotide information at positions -1 and +5. These are the two positions with the next most information after +3 and +4 (see Fig. 3). The extra information helps ($r^2 = 0.907$), but since the improvement is only slight it confirms that most of the information resides independently in the two nucleotides following the amber codon.

a)

position:		-3	-2	-1	0	+1	+2	+3	+4	+5
	A	+0.02	-0.20	+0.30		a		+1.72	-1.15	-0.22
	C	+0.35	+0.28	+0.16				+0.54	-0.98	-0.45
base	G	-0.03	+0.07	-0.53			g	+1.78	-1.55	-0.30
	T	0.00	0.00	0.00	t			0.00	0.00	0.00

+ 1.85 = ln(suppression activity)

b)



r^2 = 0.857 (corrected for degrees of freedom)

Figure 3. Six Mono-Nucleotides vs. Suppression Activity. a) Matrix for the ln(suppression activity) using the two codons surrounding the amber codon. b) Plot of observed ln(activities) versus those predicted from the matrix of part a.

Mutagenesis by 2-AminoPurine

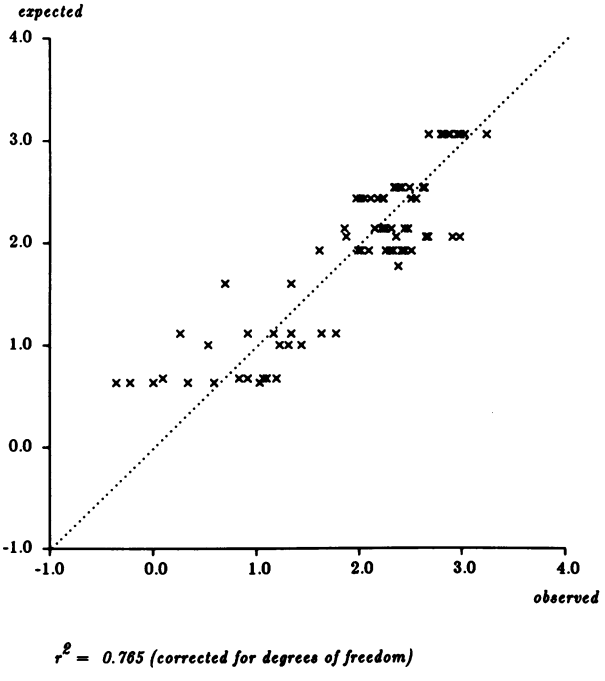
Mono-nucleotides at Positions Surrounding the Mutation. Figure 7a shows the data from Coulondre *et al.* (8), for the effect of context on mutagenesis by 2AP in the *E. coli lacI* gene. The mutation is always a **C** to **T** change, in either a **CAA** or **CAG**,

a)

position:		+3	+4
	A	+1.48	-0.82
	C	+0.33	-0.93
base	G	+1.77	-1.30
	T	0.00	0.00

+ 1.60 = ln(suppression activity)

b)



r^2 = 0.765 (corrected for degrees of freedom)

Figure 4. Two Mono-Nucleotides vs. Suppression Activity. a) Matrix for the ln(suppression activity) using the two nucleotides following the amber codon. b) Plot of observed ln(activities) versus those predicted from the matrix of part a. Note that the horizontal lines are the result of there being only 16 different predicted activities, and only 12 of the 16 di-nucleotides exist in the data set.

creating an ochre or amber mutation. Figure 7b displays the incorporation pathway that leads to the transition mutation. We first solved for the effect of the mono-nucleotides in the preceding codon plus the last base of the mutated codon (either an **A** or a **G**) and the first nucleotide of the following codon. This includes all of the

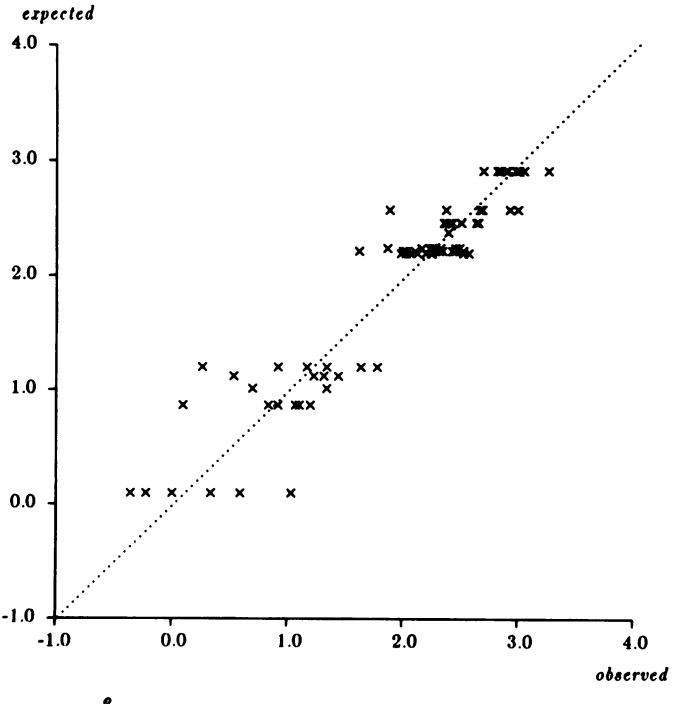
a)

di-nucleotide at positions +3,+4

		base at +4			
		A	C	G	T
base at +3	A		+1.23	+1.37	+1.90
	C	+0.19	+0.11	-0.92	+1.20
	G	+1.45	+1.19	+1.56	
	T		-0.15		0.00

+ 1.01 = ln(suppression activity)

b)



$r^2 = 0.876$ (corrected for degrees of freedom)

Figure 5. Di-Nucleotide vs. Suppression Activity. a) Matrix for the ln(suppression activity) using the di-nucleotide from positions +3,+4. Blanks in the matrix are for di-nucleotides that do not exist in the data set (see Figure 3). The di-nucleotide TT has been set to zero in this analysis. b) Plot of observed ln(activities) versus those predicted from the matrix of part a.

information in the six nucleotides that are closest to the base pair where the transition occurs, and which might influence the variation in mutagenesis. Figure 8 shows the equation and the graph for the best solution of this kind. The predictions from

a)

position:		-1	+3,+4 (di-nucleotide)			+5
			A	C	G	T	
	A	+0.22		+1.06	+1.10	+1.86	+0.30
	C	+0.09	-0.32	-0.04	-1.07	+1.09	-0.17
base	G	-0.10	+1.17	+1.05	+1.30		-0.07
	T	0.00		-0.35		0.00	0.00

\+ 1.19 = ln(suppression activity)

b)

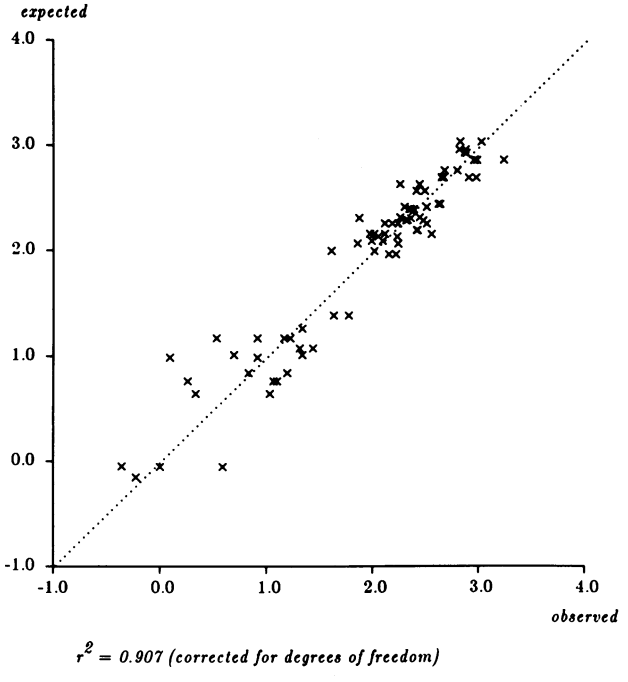


Figure 6. Di-Nucleotide + Two Mono-Nucleotides vs. Suppression Activity. a) Matrix of the ln(suppression activity) using the di-nucleotide from positions +3,+4 and the mono-nucleotides from -1 and +5. b) Plot of observed ln(activities) versus those predicted from the matrix of part a.

this model are not very good (r^2 = 0.278). The matrix weakly suggests that most of the information is in the two nucleotides at positions -2 and -1.

Di-nucleotide at the 5' Position. Figure 9 shows the equation and graph for the model that the important information is in the di-nucleotide that is 5' to the mutation. r^2 is 0.749, demonstrating that the preceding nucleotides are important, and that their effects are not independent. The activities of **CA**, **GA** and **TG** are

a)

Mutation Context							Number of Isolates	Mutation Site
-3	-2	-1	0	+1	+2	+3		
A	A	A	*c*	a	G	T	8.0	A9
A	A	A	*c*	a	G	G	9.0	A31
G	A	C	*c*	a	G	A	11.0	A16
C	A	C	*c*	a	G	C	23.0	A19
T	G	C	*c*	a	G	C	15.0	A21
C	T	C	*c*	a	G	T	13.0	A24
T	A	T	*c*	a	G	A	3.0	A5
A	T	T	*c*	a	G	C	2.0	A23
G	A	T	*c*	a	G	A	5.0	A26
T	C	T	*c*	a	G	G	7.0	A33
A	A	T	*c*	a	G	C	1.0	A35
T	C	G	*c*	a	A	A	36.9	011
G	C	G	*c*	a	A	C	39.6	017
C	A	G	*c*	a	A	A	23.4	021
A	T	G	*c*	a	A	A	69.3	029
C	T	G	*c*	a	A	C	100.8	034
G	C	A	*c*	a	A	C	72.9	035
C	A	A	*c*	a	A	C	2.7	09
C	A	A	*c*	a	A	A	9.9	010
G	A	T	*c*	a	A	C	9.9	028
A	A	T	*c*	a	A	A	4.5	013
T	T	T	*c*	a	A	C	1.8	024
G	G	G	*c*	a	A	A	4.5	027

b)

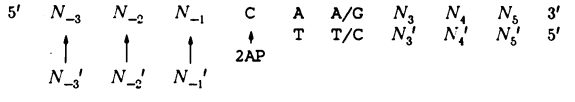


Figure 7. Context Effects on Mutagenesis. a) Each *c* at position 0 has been found mutated to a *t*, giving rise to an amber or ochre codon. The number of isolates are from Coulondre *et al.* (8). The numbers for ochres have been normalized for the difference in sample size from the amber screen. 3' to each mutation is always an a, and 3' to that is either an A or a G. Capital letter nucleotides are variables in the analysis. Three sites with the sequence CCAGG have been omitted from the analysis because the methylation of the C causes high spontaneous mutation to T (8). b) The replication fork at the point of 2AP insertion pairing with a C.

significantly higher (1.5 to 2.1) than predicted by the mono-nucleotide analysis (compare Fig. 8a to Fig. 9a). The fit to the data is still not excellent. We have not tried more complicated analyses because more data would be required to get reliable tests of more complex models.

Translational Initiation Efficiency

Mono-nucleotides at -1 Codon. Figure 10 shows the relative efficiencies of translation for one protein with 39 different triplets preceding the initiation codon (9). Figure 11 shows the analysis of these data, using the model of independent effects from those mono-nucleotides. The fit is quite poor ($r^2 = 0.249$).

a)

position:		-3	-2	-1	0	+1	+2	+3
	A	+0.02	-0.38	+1.20		a	+0.38	+0.25
	C	-0.12	+0.40	+2.17	c			+0.27
base	G	+0.29	-1.49	+2.07			0.00	+0.58
	T	0.00	0.00	0.00				0.00

\+ 0.88 = ln(number of mutants)

b)

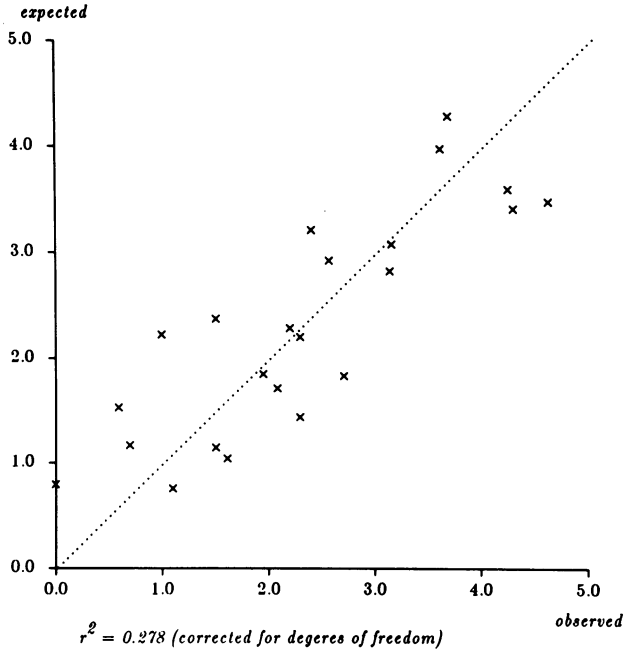


Figure 8. Mono-nucleotides vs. 2AP Mutagenesis. a) Matrix for predicting number of 2AP induced amber mutations using the surrounding six mono-nucleotides as effectors. Note that the **G** value at position +2 has been set to zero because **A** and **G** are the only possibilities there. b) The plot of observed ln(number of mutations) versus the number predicted from the matrix of part a.

Di-nucleotide at -1,-2 Plus Mono-nucleotide at -3. We tried all the di-nucleotide combinations, -3 and -2, -2 and -1, and -3 and -1, as well as combinations of each of them with each other and with the mono-nucleotides. None gave a good fit to the data; the best was the di-nucleotide at -2 and -1 plus the mono-nucleotide at -3 (r^2 = 0.299, not shown). In this example the relevant information in determining the activity is the entire triplet. It cannot be subdivided into additive parts.

a)
di-nucleotide at positions -2,-1:

-1	A	C	G	T
-2 A	+1.25	+2.13	+2.51	+0.66
C	+3.65		+3.00	+1.31
G	+2.07		+0.86	
T		+1.92	+3.79	0.00

+ 0.64 = ln(number of mutations)

b)

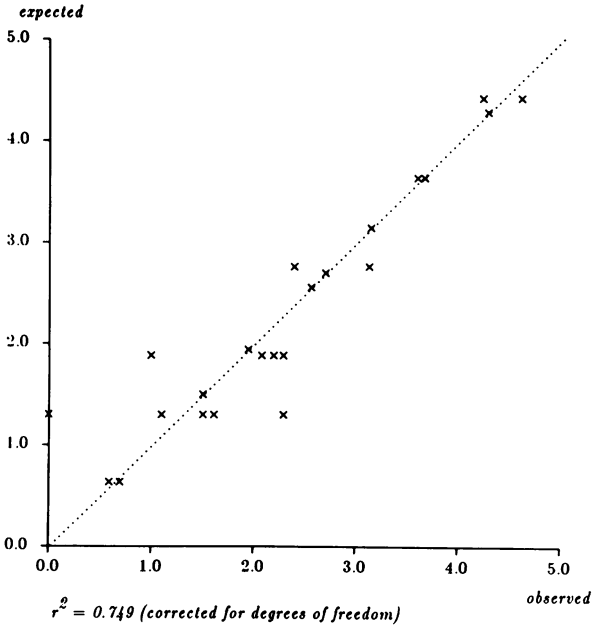


Figure 9. Di-nucleotide vs. 2AP Mutagenesis. a) Matrix for predicting the number of 2AP induced amber mutations, using the di-nucleotide preceding the mutation site. Blanks in the matrix are for di-nucleotides that do not occur in the data set. b) Plot of observed ln(number of mutations) versus the number predicted from the matrix of part a.

DISCUSSION

Suppression Activity

We found a simple description for the context effects on suppression activity. The effect is primarily localized to the two nucleotides 3' to the amber codon, and those two nucleotides affect suppression in a manner that is largely additive. A purine 3' to the amber codon increases the suppression activity about 4.5-fold rela-

context: -3 -2 -1	β-gal synthesized	context: -3 -2 -1	β-gal synthesized
T T C	0.1	A C G	0.6
T C A	0.1	A A C	0.6
A G G	0.1	A G A	0.6
T G G	0.2	G A C	0.6
C T G	0.2	C C G	0.7
A T T	0.2	T A A	0.8
T T G	0.3	C C A	0.8
C C T	0.3	C A C	0.8
C G G	0.3	C A G	0.8
G C C	0.3	C G C	0.8
G G C	0.3	T A C	1.0
T T T	0.4	T G T	1.0
T C C	0.4	C A A	1.0
G A G	0.4	A A T	1.0
G G T	0.4	A G T	1.0
T C G	0.5	G C T	1.0
T A G	0.5	G A A	1.3
A T A	0.5	T A T	2.0
A T G	0.5	C T T	2.2
G C A	0.5		

Figure 10. Context Effects on Translation Initiation. Hui *et al.* (9) varied the tri-nucleotide 5' to the initiation codon of a β-galactosidase gene. The activities are normalized to the amount made with *CAA* as the tri-nucleotide.

tive to a pyrimidine. A **T** at the next nucleotide increases suppression about 3-fold relative to any other base. The contribution of other nucleotides and the non-independent contribution of the two 3' nucleotides is about 2-fold, accounting for the total effect of about 30-fold differences in suppression activity for different contexts.

It is, perhaps, surprising that the contribution of the two bases is additive. One might expect that, since this is RNA, there will be important factors in the structure that are determined by at least pairs of bases. The data tell us that non-additive information in the di-nucleotide, and anywhere else, is small compared to the additive information of the two important mono-nucleotides. Using the matrix found by regression on just those two bases gives a fairly good fit to the data. Adding extra information only improves the fit slightly.

Suppression involves a competition between the suppressing tRNA and the release factor for the mRNA-ribosome complex (15). We cannot tell from these data which reaction is affected by the context, or whether both are. However, the nature of the context effect rules leads to a model that we find very interesting. Previous explanations of suppression enhancement by the 3' purine have relied on the observation that all tRNAs have a **T** 5' to the anti-codon (16). This could base-pair with the 3' purine to give a four base annealing, presumably strengthening the binding and improving suppression. However, Ayers and Yarus (17) have shown that the context rule is the same even if the suppressor tRNA does not have a **T** 5' to the anti-

a)

position:		-3	-2	-1
	A	+0.13	+0.95	-0.44
	C	+0.44	+0.28	-0.69
base	G	+0.07	+0.18	-0.78
	T	0.00	0.00	0.00

-0.756 = ln(relative β-gal expression)

b)

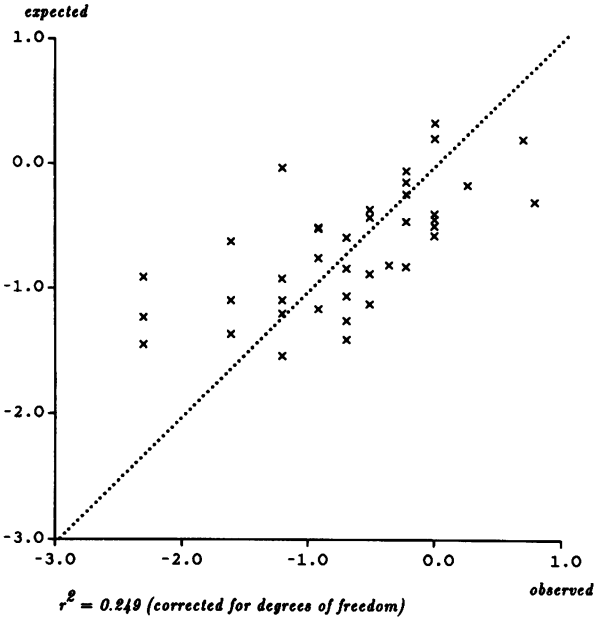


Figure 11. Mono-nucleotides vs. Expression Efficiency. a) Matrix of the relative initiation efficiencies using the mono-nucleotides from the preceding triplet. b) Plot of observed ln(β-gal made) versus that predicted from the matrix of part a.

codon; purines 3' to the amber codon lead to higher suppression even when the tRNA has a purine 5' to the anti-codon. Therefore we prefer another model that also explains the effect of the T at position +4, an effect that was not noticed in the previous analyses.

Annealing energies are enhanced by adding an extra unpaired nucleotide on the 3' end of one strand of the helix (18). This "dangling end" effect is strongest when the unpaired base is a purine. Therefore a purine following the amber codon could

increase the binding energy of the codon/anti-codon helix and increase suppression relative to release. However, there is also a force driving the +4 base, in the middle of the codon, to stack on the +3 base. If the whole two codon section of the mRNA cannot be stacked, these forces are in competition. A **T** at the +4 base, being the poorest stacker, will function to uncouple the two bases and facilitate the stacking of the +3 base. That is, a **T** at position +4 will increase the dangling end effect of the +3 base by decreasing the competing stacking interaction between bases +3 and +4. These two effects could function additively to determine the binding of the suppressor tRNA and the suppression activity.

If this model is correct, we would expect the suppression activity to be independent of the suppressor tRNA used. We have looked at the data for suppressors *su*1, *su*3 and *su*6 (not shown, data in the same papers). The result is qualitatively the same: the bases +3 and +4 are of primary importance and the rules are the same. The fit is not as good. These other suppressors are more active. Something besides the suppression activity is probably limiting the total amount of protein made.

The regression analysis has highlighted several new aspects to the effects of context on suppression. The matrix displays the strong suppression stimulation by a 3' purine, as was noticed in the original papers (6,7). In addition, the stimulation of suppression due to a **T** at position +4, not previously seen, is also evident in the matrix. The quantitative treatment has shown that the two effects are of nearly equal strength and are largely independent of each other. The effects from other nucleotides are small in comparison to the effects from positions +3 and +4.

Mutagenesis by 2-AminoPurine

In contrast to the suppression activity, the effects of the neighboring bases are not independent for the mutagenesis activity of 2-aminopurine. 2AP is an analog of the nucleotide **A** that acts as a transition mutagen (19). It can cause **G-C** to **A-T** transitions by incorporating into DNA base paired with a **C**, and then being read as an **A** in further replication. The establishment of a transition mutation by 2AP requires both the mispairing of the nucleotide analog with a **C**, and the escape of that base-pair from proof-reading by the polymerase. Mhaskar and Goodman (20) have examined both steps and found that misincorporation is independent of the base 5' to it, but that proof-reading is very sensitive to that base. Mispairs adjacent to a **G-C** base pair are repaired less frequently than those next to an **A-T** pair.

In our case study, the 2AP was always incorporated 3' to a **T** (Fig. 7b). Since the base pairs with the largest effect on that incorporation were 3' to it (N'_{-1} and N'_{-2}), they must be affecting the proof-reading. The fact that two bases are important implies that the 2AP-C mispair does not get irreversibly set in place until at least two other bases have been put in after it. This is reasonable if we assume that

the major tautomer of 2AP makes a poor base-pair with C, because two more good pairs would be needed to form a stable helix segment that would look normal to the proof-reading part of the polymerase. After those two base pairs are in place there is little effect on the mutation frequency by the next bases. The stacking free energy of the -2 and -1 base pairs (21) is not strongly correlated with their mutagenic activity. Other factors, such as rate of polymerase incorporation, must influence the proof-reading efficiency.

The lack of data is limiting in this example. It would be interesting to see the values for the missing elements of Fig. 9. The missing CC is predicted from the mono-nucleotide analysis (Fig. 8) to have the highest value. In addition, more data would allow testing of more complex models. The 23 data points (Fig. 7) means we can solve for at most 23 variables, but reliable results usually require many more data points than unknowns (i.e. high degrees of freedom). In the analyses of Figures 8 and 9 we have solved for 14 and 12 unknowns, respectively, leaving only 9 and 11 degrees of freedom. To add more complexity to the analysis of Figure 9, by including more positions as variables, would degrade the reliability unless more data were available. We include this example to show how analyses more complicated than mono-nucleotides can be performed by the regression method. The appropriateness of the model can be determined by comparing the observed values with those predicted from the analysis. In the suppression case we learned that little was gained by using models more complicated than mono-nucleotides. In this case di-nucleotides are much superior but still not completely adequate.

Translational Initiation Efficiency

In this example neither mono- nor di-nucleotide effects could be added together to predict the activity of the tri-nucleotide. The information in the whole triplet is important to initiation. We include this example not to undermine the importance of the regression method, but rather to show that even negative results give valuable information. What we learn in this analysis is that the entire triplet, as a whole, is important in determining the translational efficiency. Any quantitative models about ribosomes binding to these sites must include that information to be reliable. This limits the type of models that one need consider.

The sequence surrounding the translation initiation site in this construct is:

5'- tcacac**AGGA**aacagaattct**GGAGG**tctagnnn***ATG***T**GT**Gatctggatccc -3'.

The **ATG** 3' to the nnn is the initiation codon for the β-galactosidase protein. Adjacent, and overlapping, are two out-of-frame **GTG** codons that can serve as initiation codons (22-24). The **ATG** would be expected to be used as the initiation codon most often (24,25), but the competing **GTG** codons might direct some ribosomes into the wrong frames (26). The variable nucleotides might have some influence on the rela-

tive use of the three potential initiation sites. There are also two Shine-Dalgarno sequences 5' to the coding region. Near the 5' end is the sequence **AGGA**, and nearer the **ATG** is the sequence **GGAGG**. Only the latter has the proper spacing from the initiation codon to give high level translation (22-24). However, we have recently discovered that an upstream Shine-Dalgarno sequence can bind ribosomes and inhibit synthesis from a normal initiation site (Barrick *et al.*, in preparation). In addition, there are two potential secondary structures that occur within this region. The bases **TCACACA**, at the 5' end of the sequence, are complementary to the **TGTGTGA** of the overlapping initiation codons. Some variable region sequences could extend this potential structure. The **GGAGGTCTAG**, from the Shine-Dalgarno sequence to the variable nucleotides, is complementary to the **CTGGATCCC** at the 3' end of the sequence shown. This potential secondary structure can also be extended by some of the variable region sequences. We are left with a very complicated picture of the possible interactions of this sequence with ribosomes. Besides direct effects of the variable nucleotides on the efficiency of using the **ATG**, there are going to be effects on which structures are more favored, which Shine-Dalgarno sequence is better bound and how efficiently the different initiation codons are used. In retrospect we are not surprised that the variable nucleotide effects are complex.

CONCLUSIONS

Determining Sequence/Activity Relationships

Matrices provide a convenient way to evaluate nucleic acid sequences. When activities are known for a collection of sequences, that information can be used to construct matrices that evaluate sequences according to their activities. Whenever the activity of a sequence can be decomposed into parts that add together to give the total activity, the matrix defined by multiple regression will have predictive value. In some cases the important information is in mono-nucleotides, as in the suppression example. In the mutagenesis example the important information lies in the di-nucleotide preceding the mutation spot. For the initiation efficiency data of Hui *et al.* (9), the fit for all models (short of the entire triplet as the unit of information) was not very good. The regression method not only finds an evaluation for a sequence's activity, but also tells one the units of information that are important.

Alternative Matrices

In each of the examples of this paper there is an obvious alignment of the sequences. For some biological recognition problems there will not be an obvious alignment. For *E. coli* promoters, there are (at least) two regions of recognition and the spacing between them is variable (14). One can still use the matrix evaluation method, as did Mulligan *et al.* (5), by aligning on the two regions independently and

adding another term for the spacing variable. They used a function of the nucleotide frequencies in promoter sequences as the matrix elements and obtained fairly good correlation between promoter evaluations and measured activities. More data would make it possible to solve for the matrix which gave the best fit to the data.

For some problems it may not be obvious what is recognized or how to align the sequences. Galas *et al.* (27) have presented a method to find important patterns in sequences with only approximate alignments. They look for short oligo-nucleotides that are unusually common, including closely related oligos. These are found within "windows", short regions about the approximate alignment, so that the exact positions of the oligos are not important. We have previously described a program that will flexibly encode sequences into numeric vectors (12). Those vectors simplify the multiple regression analysis described in this paper. It is also straightforward to extend the vectors along the lines of Galas *et al.* (27) to deal with problems of variable or imprecise alignment of sequences.

Acknowledgements. We thank Mike Yarus and Doug Turner for helpful discussions. This work was supported by NIH grant GM28755. Computer resources were from NIH grant RR01538 and the University of Colorado Academic Computing Center.

References

1. Stormo, G.D., Schneider, T.D., Gold, L. and Ehrenfeucht, A. (1982) Nucl. Acid Res. 10, 2997-3011
2. Harr, R., Haggstrom, M. and Gustafsson, P. (1983) Nucl. Acid Res. 11, 2943-2957
3. Staden, R. (1984) Nucl. Acid Res. 12, 505-519
4. Brendel, V. and Trifonov, E.N. (1984) Nucl. Acid Res. 12, 4411-4427
5. Mulligan, M.E., Hawley, D.K., Entriken, R. and McClure, W.R. (1984) Nucl. Acid Res. 12, 789-800
6. Miller, J.H. and Albertini, A.M. (1983) J. Mol. Biol. 164, 59-71
7. Bossi, L. (1983) J. Mol. Biol. 164, 73-87
8. Coulondre, C., Miller, J.H., Farabaugh, P.J. and Gilbert, W. (1978) Nature 274, 775-780
9. Hui, A., Hayflick, J., Dinkelspiel, K. and de Boer, H.A. (1984) EMBO J. 3, 623-629
10. Ryan, T.A. Jr. (1981) Minitab Student Handbook, Duxbury Press, Boston MA
11. Schneider, T.D., Stormo, G.D., Haemer, J.S. and Gold, L. (1982) Nucl. Acid Res. 10, 3013-3024
12. Schneider, T.D., Stormo, G.D., Yarus, M.A. and Gold, L. (1984) Nucl. Acid Res. 12, 129-140
13. Edwards, A.L. (1979) Multiple Regression and Analysis of Variance and Covariance, W.H. Freeman and Co., San Francisco
14. McClure, W.R. (1985) Ann. Rev. Biochem. 54, 171-204
15. Caskey, C.T., Forrester, W.C. and Tate, W. (1984) in Gene Expression, Alfred Benzon Symposium 19, Clark, B.F.C., and Peterson, H.U., eds., Munksgaard, Copenhagen, pp. 149-158

16. Sprinzl, M. and Gauss, D.H. (1984) Nucl. Acid Res. 11, r1-r57
17. Ayer, D. and Yarus, M. (1986) Science 231, 393-395
18. Freier, S.M., Burger, B.J., Alkema, D., Neilson, T. and Turner, D.H. (1983) Biochem. 22, 6198-6206
19. Ronen, A. (1979) Mutat. Res. 75, 1-47
20. Mhaskar, D.N. and Goodman, M.F. (1984) J. Biol. Chem. 259, 11713-11717
21. Tinoco, I., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. and Gralla, J. (1973) Nature New Biol. 249, 40-41
22. Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B.S. and Stormo, G. (1981) Ann. Rev. Microbiol. 35, 365-403
23. Stormo, G.D., Schneider, T.D. and Gold, L. (1982) Nucl. Acid Res. 10, 2971-2996
24. Stormo, G.D. (1986) Translation Initiation, in Maximzing Gene Expression, Reznikoff, W. and Gold, L. eds., Benjamin/Cummings Publishing Co., Inc., pp. 195-224.
25. Childs, J., Villanueba, K., Barrick, D., Schneider, T.D., Stormo, G., Gold, L., Leitner, M. and Caruthers, M. (1985) in Sequence Specificity in Transcription and Translation, Alan R. Liss Publishing, pp. 341-350
26. Wulff, D.L., Mahoney, M., Shatzman, A. and Rosenberg, M. (1984) Proc. Natl. Acad. Sci. USA 81, 555-559
27. Galas, D.J., Eggert, M. and Waterman, M.S. (1985) J. Mol. Biol. 186,117-128